DOCUMENT RESUME

ED 300 851 CS 506 463

AUTHOR van Gelderen, A.
TITLE Differential Rating of Performance on Oral Tasks in a Large Scale Survey in the Netherlands.
PUB DATE Apr 87
NOTE 19p.; Paper presented at the International Oracy Convention (Norwich, England, April 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Communication Research; *Construct Validity; Elementary Education; Foreign Countries; Models; Reliability; *Speech Communication
IDENTIFIERS Netherlands

ABSTRACT

At the Educational Research Centre (S.C.O.) in Amsterdam, a study determined the applicability and construct validity of ratings of speaking performances by examining tape-recordings of subjects in four dimensions. Subjects were 200 pupils of 11 and 12 years of age, and performances on four different oral tasks were investigated. The rating dimensions--defined as main functions of speaking--were reference, delivery, fluency and articulacy. Each main function was represented by a rating category that was selected in a pretest. Inter-rater reliabilities, correlations between panel ratings--within and between tasks--were calculated and some hypothetical models for correlations between ratings were tested. Results indicated that performances on two oral tasks were rated on the four dimensions in a meaningful way. Performances on the other tasks were rated in such a way that only reference and articulacy were meaningfully differentiated. Furthermore, ratings on articulacy correlated highly between tasks, whereas ratings on reference correlated relatively weakly between tasks. (Five tables of data are included, and 12 references are attached.) (RAE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED300851

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
A.J.S. van Gelderen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Differential Rating of Performances on Oral Tasks in a Large Scale Survey in the Netherlands

A. van Gelderen

Paper presented at the International Oracy Convention at Norwich (April 1987)

Summary

This paper describes research carried out at the Educational Research Centre (S.C.O.) in Amsterdam to determine the applicability and construct validity of ratings of speaking performances from tape-recordings of 200 pupils of 11 to 12 years of age in four dimensions. Performances on four different oral tasks were investigated. The rating dimensions — defined as main functions of speaking — were Reference, Delivery, Fluency and Articulacy. Each main function was represented by a rating category that was selected in a pretest. Inter-rater reliabilities, correlations between panel ratings — within and between tasks — were calculated and some hypothetical models for correlations between ratings were tested. Results indicate that performances on two oral tasks were rated on the four dimensions in a meaningful way. Performances on the other tasks were rated in such a way that only Reference and Articulacy were meaningfully differentiated. Furthermore, ratings on Articulacy correlated highly between tasks, whereas ratings on Reference correlated relatively weakly between tasks. Implications of the findings are discussed in terms of taxonomic task-characteristics and their relation with categories that are used to evaluate performances on those tasks.

CS506463

1. Introduction

In a national survey in the Netherlands in 1985 performances of 200 pupils of 11 to 12 years of age on several functional oral tasks were collected in order to determine their level

of spoken language. Performances on three tasks were rated by a jury using the following rating categories:

- Language usage, in which a composite judgement is realised of the quality of word-choise, sentence-structure, articulation and tempo.
- Organisation, meaning the extent to which a logical sequence of main points and details is realised by the speakers.

Furthermore each performace on these tasks was given a score for Content, indicating the total of content-elements that were mentioned by the speaker.

In the contributing paper of Van den Bergh to this conference (1987) these ratings and this score are described in more detail. Information regarding the oral tasks on which the performaces were gathered is to be found in the same paper. We are now considering the rating of performances on the following tasks: Retelling a story (task 1), Reporting an accident (task 2) and Explaining how a spider constructs his web (task 3).

A problem emerged in determining the informative value of the ratings and scores on the performances in the light of the high intercorrelations between these ratings within tasks. In tabel 1 these intercorrelations of jury-ratings and scores per task are given.

Tabel 1: Intercorrelations between jury-ratings on Content, Language usage and Organisation on three oral tasks (above diagonals) and after correction for attenuation (below diagonals); on the diagonals: the inter-rater-reliabilitys (alpha's), N=200

	Task 1 (story)			Task 2 (accident)			Task 3 (spider)		
	CON	ORG	L.V.	CON	ORG	L.V.	CON	ORG	L.V.
CON	(.98)	.82	.71	(.95)	.74	.64	(.95)	.83	.79
ORG	.86	(.93)	.83	.80	(.89)	.90	.88	(.93)	.91
L.U.	.76	.91	(.89)	.70	1.01	(.89)	.84	.98	(.92)

The correlations after correction for attenuation in Table 1 (below the diagonals) are estimates of the true correlations between the ratings when unreliability due to rater differences is taken into account. The correlations between the ratings on Organisation and Usage within tasks are extremely high and make them hardly discriminable in a psychometric sense. The correlations between these ratings and the scores on Content are somewhat weaker, although they still constitute indications of large proportions of shared variance.

It was hypothesized that the concepts of Organisation and Usage were not sufficiently well-defined to be used as rating categories for an analytical differentiation between qualities of speaking performances. From research into rating procedures it is a well-known fact that a lack of conceptual clarity of rating categories gives rise to undesirable influences on raters, such as the halo-effect (Saal et al., 1980).

A follow-up study was therefore conducted to find out whether a scheme of rating categories could be found that would allow a more differential interpretation of the speaking performances.

In this paper theoretical considerations and data analyses of this follow-up study will be given that are relevant to the following questions.

1. Can a rating scheme be found that is applicable to the different oral tasks?
2. What are the relations between the elements of that scheme when applied to the different oral tasks?

First the theoretical background for the development of the rating scheme will be discussed and the — so called — main functions that the scheme comprises are defined. Next a systematic comparison between the three tasks is given in taxonomic terms. From this it is concluded that the three tasks fail to vary on one important taxonomical dimension, namely thematic freedom. For this reason performances on a fourth task were added for a comparison with the other tasks. Subsequently the selection of rating categories for each task and the results of a pretest are briefly reported. The final selection of rating categories is presented and the rating

procedures are described. The results are presented in terms of inter-rater reliabilities and correlations of panel ratings within and between tasks. Furthermore some additional data concerning construct validity of the panel ratings per task are given. Finally the results are discussed, comparing the application of the 'old' rating scheme (presented above) with that of the 'new' one and concentrating on some findings that resulted from the application of the 'new' scheme to performance on four oral tasks.

2. Theoretical background of the new rating scheme

The rating scheme we developed is based partly on Becker's research (1962) on the rating of speeches. For the other part we extended the scope of application beyond the formal speech situation by using a functional perspective in the conceptualisation of rating dimensions. This, we believe, can be based upon Crystal and Davy's distinction (1979) between Fluency, Intelligibility and Appropriateness as main factors of success in conversational speech.

Becker (1962) found three dimensions in which specific rating categories for the rating of speeches could be divided. He based his conclusion on the fact that different rating categories from one dimension were not sufficiently discriminated in a psychometric sense:

> 'The goal should be to have each scale measure a relatively independent aspect of the performance. Perhaps the speech form should be reduced to three scales, a content-analysis scale, a delivery scale, and a language scale'. (Becker, 1962, o.c. 44).

This distinction between content-analysis, delivery and usage as a sensible rating scheme has been widely adopted in research on rating of oral performances. Often, however, the scheme is adjusted to fit in more precisely with the demands of the speaking tasks. So we have found a diversity of formulations of rating categories that would fit into one of the dimensions of Becker (e.g. Wesdorp, 1981; Hitchman, 1865, Rijlaarsdam & Bronkhorst, 1983).

In developing the new rating scheme we reformulated Becker's dimensions in order to make them applicable to the performances on functional oral tasks. We generalized the dimensions in such a way that they could be regarded as main functions that speakers have to perform in order to be clearly understood by a listener in all kinds of communicative situations. In doing so the following four main functions emerged:

1. Reference. This function is a generalisation of Becker's 'content-analysis' dimension. The term has been borrowed from Dickson's work on referential communication (e.g. 1982 a and b). It pertains to the informational value of the speaking performance to the listener. This is determined by the clarity of the content elements (to the listener) that occur in the performance and by the sequence and structure in which they occur. This function encompasses categories like 'Organisation', 'Relevance', 'Clarity of ideas' and 'Significance'.
2. Delivery. This function is virtually identical to Becker's 'delivery' dimension when articulation and pronunciation are excluded from the definition. The delivery function is also expressed in what Crystal and Davy (1979) termed appropriateness and defined as 'the suitability of language to the situation'. In all communicative situations speakers ought to behave in a way that listeners perceive as being in accordance with subject and context. By means of posture, attitude, paralinguistic features of speech (intonation patterns, volume-variation) or vocabulary that is used to make a specific impression, a speaker determines the degree and kind of the listener's involvement.
3. Fluency. This function is identical to the main factor of conversational speech that Crystal and Davy termed fluency. Becker differentiated between 'language' (as a dimension) and 'fluency' (as a rating category for delivery) but the reason for his doing so is conceptually unclear. Fluency is defined by the impression of smoothness of continuity in discourse. This impression is determined by the variety of sentence patterns, the way

sentences are connected and the ease with which words are found that fit into the flow of speech. In general a fluent speaker optimizes interpretability by using periodic, linguistic and prosodic cues that indicate where a sentence ends and where a new one begins. Moverover, in speaking performances of a relatively long duration, these cues may indicate where a passage ends in order to be followed by a new one and where a passage ends that is intended to be a concluding one.

4. Articulacy. This function is identical to Crystal and Davy's main factor, termed 'intelligibility', which they defined as the recognizability of the words and sentence patterns of speech. It involves the realisation of phonetic characteristics of the vowel and consonant system in speech. Becker did not distinguish these articulatory phenomena from his delivery dimension. Conceptually, however, there is a clear distinction between the recognizability function and the involvement-effecting function of speech. The first is in a way a prerequisite of the last.

3. A systematic comparison of the oral tasks

In this section I shall consider some important characteristics of the oral tasks in more detail. This is necessary because such characteristics might have a profound influence on the application of rating categories that are used to evaluate the performances on these tasks. For example, Wilkinson and Stratta (1965) contend that the relevance of rating categories -- which they divide into Sound, Load and Transport -- depends on the kind of speaking situation.

> 'High Load, inadequate Transport, would result in failure to communicate: low Load, high Transport, would be inefficient in intellectual discourse, but in affective or phatic usages probably essential ...'. (Wilkinson & Stratta, 1969, o.c. 186).

What then, are the distinctive features (and similarities) in our oral tasks that could affect the application of rating categories? I distinguished four taxonomic dimensions of oral tasks that are of theoretical relevance, as follows:

- The functional context that defines a speaker's general purpose. Such purposes are amusing, informing, expressing, ritualising, discursive, convincing, manipulating or explaining. The functional context is not described in terms of an act (like 'describing') but in terms of the intention that is hidden in the act (like 'informing').
- The speaking mode that categorizes the speaking situation in terms of global characterisations of concrete acts. Typical speaking modes are telling a story, giving a description, a report or a speech, conversing, discussing and so on.
- The thematic structure that defines the degree of freedom speakers have in determining the subject of their talk and the way they elaborate it. Formal speaking situations are characterised by a narrowly defined structure and an accompanying low degree of thematic freedom. Informal situations, such as conversations and talks, are characterised by a relatively loosely defined structure and a high degree of thematic freedom.
- The interactive structure that defines the degree of interaction that is allowed in the speaking situation between 'speakers' and 'listeners'. We can distinguish between three interactive structures: the monologue, the dialogue and the polylogue. In the monologue the listener is not supposed to interrupt for a considerable amount of time. In the dialogue he takes turns with the speaker, and in the polylogue many can take turns in contributing to the communication.

When we compare the three oral tasks using these taxonomic dimensions the following picture emerges.

Table 2: Three oral tasks, described in terms of functional context, speaking mode, thematic freedom and interaction structure

	Functional Context	Speaking Mode	Thematic Freedom	Interaction Structure
Task 1 (retelling a story)	Amusing	Story	Restricted	Monologue
Task 2 (reporting an accident)	Informing	Description	Restricted	Dialogue
Task 3 (explaining web construction)	Explaining	Exposition	Restricted	Monologue

It will be seen in Table 2 that the three tasks considered vary in all but one of the four taxonomic dimensions. In the dimension of thematic structure they show virtually no variation. Variation in this dimension seems important, though, in the context of an investigation in which the applicability of a general rating scheme is tested. Such a rating scheme should also be applicable in evaluating speaking performances where speakers are relatively free to determine their subject.

3.1. Adding a fourth task

For this reason a fourth oral task was developed to represent speaking situations in which the speaker is allowed a considerable degree of thematic freedom. A detailed description of this fourth task — completing a story — cannot be given here. It has many features in common with task 1 (retelling a story), with the exceptions that the pupils only hear a part of the story and that they do not have to retell that part, but have to complete it using their own imagination and relate it to a fellow pupil.

This task, together with task 1, was given to 100 pupils from twelve elementary schools, and the performances were tape-recorded for the purpose of applying the rating scheme and comparing the results with those on the performances on the other tasks. More detailed information concerning this task and the performances that were collected is given in Van Gelderen (1987).

4. The selection and pretesting of rating categories

This section gives a brief description of the selection of rating categories for the evaluation of the performances on each of the four oral tasks. Detailed information of the elaborate pretesting procedures is given in Van Gelderen (1986 and 1987).

First a selection was made of all the rating categories that were found in the literature on speech evaluation on the basis of face-validity for the performances of the oral tasks and mere technical applicability. (Examples of rating categories that where regarded as inappplicable in a technical sense are bodily posture and eye contact, because of the absence of visual information and volume due to the volume of the tape-recordings not having been standardized.) Furthermore, care was taken to select categories that could be brought under the heading of one of the four main functions of our rating scheme. This resulted in the selection of at least two rating categories per main function per task. An exception to this rule was Articulacy; for this main function only one appropriate category per task was found, namely articulation.

Next, this selection was printed on rating forms and given to a panel of 4 to 5 raters. The panel received intensive instruction in which the rating categories were defined, both orally and in written form, and applied the categories to performances that were selected as examples of good, mediocre and poor speaking on each task. Subsequently they scored 40 performances per task while listening to each performance

twice. This procedure was repeated for the rating of performances on each task.

On the basis of these ratings, inter-rater reliabilities and intercorrelations between panel ratings were calculated for each rating category per task. In general, the results indicated very strong relations between rating categories for the same main functions per task and weaker relations between rating categories for different main functions from the rating scheme. In most cases rating categories within the same main functions per task were indiscriminable in a psychometric sense, corroborating Becker's findings (1962). Articulacy, however, which in Becker's results was not discriminated from categories for Delivery, now showed moderate correlations with all the other main functions.

These results justified a final selection of one rating category per main function per task. For task 2, however, rating categories for Fluency and Delivery failed to meet the criteria for selection. In these cases categories were nevertheless selected to gain more certainty concerning the applicability or inapplicability of these categories to rating performances on this task.

The selected categories per task and main function that were used for the final rating of the speaking performances on the four tasks are presented in table 3 below.

Table 3: The final selection of rating categories per task and main function

Main function	Task 1	Task 2	Task 3	Task 4
Reference	relevance	relevance	relevance	relevance
Delivery	speaking vein	speaking vein	intonation	speaking vein
Fluency	pace	sentence structure	sentence structure	pace
Articulacy	articulation	articulation	articulation	articulation

5. Rating procedures

A panel consisting of three (female) raters rated 200 performances on tasks 1, 2 and 3 and another 100 performances on tasks 1 and 4, after receiving instructions that were virtually indentical to those described in the last section.

Rating categories were scored on 5-point scales. Both ends of the scales and the centre were briefly defined. The scales for Reference, for example, stressed the importance of a clear distinction between main events and details in the story-telling tasks and of a representation of main content elements in the other two tasks (there were no details to be told in the latter). The definitions of the Delivery scales were more task-specific because of the strong influence of context on the appropriateness of speech in these tasks. For the two story-telling tasks the importance of vividness was stressed; for the reporting task (task 2), however, it was mainly seriousness that had to be rated; the performances on the exposition task were rated on variety of intonation.

6. Results

In this section inter-rater reliabilities and intercorrelations between panel ratings are presented: the implications will be discussed in terms of their relevance to the two questions that were mentioned in the introductory section.

The first question concerned the applicability of the rating scheme for evaluation of the performances on the four tasks. The answer to this question is mainly determined by the psychometric discriminability of the panel ratings within tasks.

The second question concerned the relationships that exist between the elements of the rating scheme across tasks. These relationships can be derived from the correlations of panel-ratings between tasks. Of course, the answers to these questions are not independent of considerations of construct validity. For this reason the remainder of this section is

devoted to indicators of construct validity that were gathered in separate analyses that cannot be considered in detail here.

6.1. Correlations of panel ratings within and between tasks

In table 4 the correlations between panel-ratings on rating categories for task 1 to 3 are presented. Subsequently, in table 5 these correlations are presented for task 1 and task 4, which were examined separately. In these tables the names of the rating categories are replaced by the abbreviated names of the main functions they represent. On the diagonals the inter-rater reliablities are given (Cronbach's alpha, considering the scores of each individual rater as the scores on a test item). Correlations between panel ratings on categories for the same main function are underlined.

Table 4: Correlations between panel-ratings on rating categories for tasks 1 to 3 (above diagonal) and corrections for attenuation (below diagonal); on the diagonal: inter-rater reliabilities (alphas); N=200

	Task 1 (retelling a story)				Task 2 (accident)				Task 3 (spider)			
	REF	DEL	FLU	ART	REF	DEL	FLU	ART	REF	DEL	FLU	ART
REF	(.91)	.72	.54	.47	.35	.42	.40	.39	.50	.49	.40	.40
DEL	.80	(.89)	.67	.60	.30	.46	.47	.55	.43	.65	.50	.54
FLU	.62	.78	(.82)	.56	.32	.40	.52	.48	.35	.45	.43	.41
ART	.54	.69	.67	(.84)	.19	.34	.39	.69	.33	.53	.47	.70
REF	.39	.34	.38	.33	(.87)	.48	.66	.29	.24	.16	.24	.19
DEL	.51	.56	.51	.43	.59	(.75)	.62	.44	.31	.38	.36	.39
FLU	.48	.57	.65	.48	.81	.82	(.77)	.51	.38	.41	.43	.41
ART	.47	.66	.60	.86	.35	.58	.66	(.77)	.36	.48	.44	.66
REF	.57	.49	.42	.39	.28	.39	.47	.44	(.85)	.53	.72	.44
DEL	.57	.77	.56	.65	.19	.49	.52	.61	.64	(.80)	.66	.65
FLU	.49	.62	.55	.60	.30	.48	.66	.58	.91	.86	(.74)	.63
ART	.46	.63	.50	.84	.22	.49	.57	.83	.52	.80	.80	(.83)

Table 5: Correlations between panel-ratings on rating categories for task 1 and task 4 (above diagonal) and corrections for attenuation (below diagonal); on the diagonal: inter-rater reliabilities (alphas); N=100

	Task 1 (retelling story)				Task 4 (completing story)			
	REF	DEL	FLU	ART	REF	DEL	FLU	ART
REF	(.91)	.61	.53	.36	.30	.27	.26	.36
DEL	.67	(.90)	.63	.29	.33	.63	.41	.36
FLU	.60	.71	(.87)	.42	.20	.39	.53	.37
ART	.40	.32	.48	(.88)	.18	.29	.34	.76
REF	.34	.37	.23	.21	(.87)	.58	.42	.25
DEL	.30	.72	.45	.33	.67	(.86)	.66	.38
FLU	.29	.47	.61	.39	.48	.77	(.86)	.41
ART	.41	.41	.43	.87	.29	.44	.48	(.86)

Consider the above tables as consisting of 9 (table 4) and 4 (table 5) matrices of 4 by 4 correlations. This will facilitate interpretation of their meaning. The three matrices along the main diagonal of table 4 consist of the correlations between panel ratings within tasks. The same holds for the two matrices along the diagonal in table 5. The other matrices consist of correlations between panel ratings between tasks, corrected or not corrected for attenuation.

There are four major rules that apply to the correlations in both tables; they are the following:

a) For each pair of rating categories between tasks there is a category (not a member of that pair) that has a higher correlation with a member of that pair within the same task.

b) The exception to rule a are the panel ratings on Articulacy; all pairs of rating categories for Articulacy show higher correlations than any of these categories with each other rating category.

c) Correlations between panel ratings for Reference between tasks are relatively low. (Corrected for attenuation .39, .57, .28 and .34 respectively.)

d) All correlations are sufficiently low to discriminate between panel ratings in a psychometric sense.

Furthermore there is one major difference between the correlations in table 4 as compared with those in table 5. This has to do with the pattern of correlations within tasks. In table 5 these patterns are virtually the same; that is to say: pairs of rating categories that have high correlations within task 1 also have high correlations in task 4 and vice versa; the same holds for low correlations between rating categories in these tasks.

In table 4, however, each task has a quite different pattern of correlations between rating categories. It seems that performances on task 1 and task 4, despite the difference in freedom of thematic structure, have a lot in common in terms of our rating scheme.

In short, our first question has to be answered in the affirmative on the basis of these results: the rating scheme is applicable for the evaluation of the performances on the four tasks. At least, it results in panel ratings that are psychometrically distinct. The second question, however, seems to demand a more complex answer. The results of applying the rating scheme depend on the oral tasks on which performances are evaluated. This can be concluded from the relatively low correlations between rating categories for the same main function across tasks (excepting those for Articulacy). (See rules a and b above.) It can also be concluded from the different patterning of correlations within tasks in table 4.

6.2. Some indicators of construct validity

Several subsequent analyses have been performed to gain additional insight into the construct validity of the panel ratings that cannot be reported in detail here. These include the testing of models to account for the correlations between the scores of each individual rater, using Lisrel analysis, and the computing of correlations between the panel ratings for Reference with the scores for Content that were described

in the introductory section. These scores can be regarded as an external criterion for the ratings of Reference, although they are not supposed to measure exactly the same thing.

To start with, the correlations between scores for Content and ratings for Reference for tasks 1 to 3 were high enough to support the validity of the ratings; the correlations were (corrected for attenuation) .94, .63 and .87 respectively. None of the other rating categories correlated with the scores nearly as highly.

The testing of models did not result in an explanatory model that fitted the correlation matrix for the scores of the three individual raters on all tasks. It did reveal, however, that a considerable amount of surplus correlation existed among the rated categories of tasks 2 and 3 (as compared with those of task 1 and 4). Moreover, inspection of the correlations between the scores of individual raters for these tasks made it apparent that problems existed on the level of inter-rater reliabilities that remained undetected at the level of panel ratings. For tasks 2 and 3 it was not exceptional for higher correlations to exist among <u>different</u> rating categories scored by the same rater than among the scores of two raters rating the <u>same</u> category. In fact this was the case in 5 out of 18 possible instances in task 2 and in 10 out of 18 in task 3. This phenomenon did not occur in task 1, and it occurred only once in task 4.

These results are a clear refutation of assumptions of a valid application of the rating categories for performances on task 2 and 3. This is particularly true of the ratings for Fluency and Delivery in these tasks, because there is much less reason to doubt the validity of the ratings for Reference and Articulacy. Reference ratings are supported by the external criterion provided by the scores on Content. Ratings for Articulacy, on the other hand, proved to be highly stable across tasks, as can be seen in tables 4 and 5.

7. Discussion

When we compare the results presented above with those of the original rating that we described in the introductory section and presented in table 1, the following interpretation seems reasonable. Using general categories — with identical formulations across tasks — to rate performances on different oral tasks has the effect of blurring the conceptual boundaries between these categories. So, when such general rating categories are applied to the performances on one task — as has been shown in the introductory section in the case of the categories 'Usage' and 'Organisation' — the resulting ratings will have hardly any differential value. If, however, categories are derived from a general scheme and subsequenty translated in task-specific terms when needed[1] — as has been done in the study reported here — the conceptual boundaries between the derived categories become clearer. Applying these categories to performance on one task leads to better discrimination between aspects of the performances relevant to each of the categories.

Even then, however, problems in the application of rating categories have been shown to exist. More specifically, not all the tasks investigated are equally suited to an application of the rating categories that were derived from the same rating scheme. Performances on two tasks — 'accident' (task 2) and 'spider' (task 3) — were rated in a way that casts serious doubts on the construct validity of Delivery and Fluency scores. On the basis of this study it is not possible to tell what features of these tasks are responsible for these phenomena. Is it the dialogical nature of task 2 by which the contribution of the reporting pupil is hard to separate from the questions and answers of the 'police'? Or is it the strictly informative nature of the speaking situation by which the functions of Delivery and Fluency

---

1 Of course, not all rating categories have to be put in task-specific terms; e.g. categories for Articulacy should be formulated in a task-independent way.

become less important or even superfluous? Is it the explanatory nature of task 3 by which it is hard to qualify differences in the sentence structures used by the pupils? These kinds of questions will have to be addressed in subsequent research in which such taxonomic differences between tasks are systematically varied.

Another perspective on the results of this study is given when we concentrate on the two tasks for which the rating scheme did function as was intended. From this result it can be concluded that the degree of thematic freedom the speakers have — which is the main feature that distinguishes both tasks — has no significant influence on the applicability of the rating scheme. On the other hand, we still do not know which of the common features of both tasks is of most importance for the application of the scheme. Is it their narrative nature, their amusing function, or the fact that both are extended monologues? (The 'spider' task also consists of a monologue but it can hardly be called 'extended' in view of the small amount of time the pupil talks.)

Gaining more insight into the relation between task characteristics and the categories with which the performances on these tasks are sensibly and validly evaluated is not only of interest to large-scale oracy surveys. Results of such taxonomic research can also be of help to educational specialists in selecting and constructing oral tasks that are specifically designed to make pupils aware of distinct phenomena of speech. This applies to the selection of tasks in which specific main functions of speech are highlighted. It applies also to the realisation of speech-functions that are mastered in a more task-specific way (e.g. Reference and Delivery) and those that are fairly independent of task (e.g. Articulacy).

Literature

Becker, S.L. (1962). The rating of speeches: scale independence. Speech Monographs. 29, 38-44.

Bergh, H., van den. (1987. Large Scale Oracy Assessment in the Netherlands. Paper presented at the International Oracy Convention at Norwich, April 1987.

Crystal, D. & Davy, D. (1979). Advanced conversational english. London: Longman.

Dickson, W.P. (1982). Two decades of referential communication research: a review and meta-analysis. In C.J. Brainerd & M. Pressley (Eds.). Verbal processes in children (pp 1-33). New York: Springer-Verlag.

Dickson, W.P. (1982). Creating Communication-Rich Classrooms: Insights from the socioliguistic and referential traditions. In L.C. Wilkinson (Ed.), Communicaling in the classroom. (pp 131-146). New York: Academic Press.

Gelderen, A. van (1986). De validatie van analytische beoordelingen van spreekprestaties. Tijdschrift voor Taalbeheersing, 8, (3), 204-221.

Gelderen, A. van (1987). Taalmaten; Constructie van gedetailleerde beoordelingsprocedures voor spreken en schrijven ten behoeve van peilingsonderzoek. Deel 1: Het beoordelen van spreekprestaties. (S.C.O.-rapport) Amsterdam: S.C.O.

Hitchman, P.J. (1965). The testing of spoken english: a review of research. Educational Research, 7, 55-72.

Rijlaarsdam, G.C.W. & Bronkhorst, H. (1983). Beoordelen van spreekbeurten. (S.C.O.-rapport) Amsterdam: S.C.O.

Saal, F.E., Downey, R.G. & Lahey, M.A. (1980). Rating the ratings: assessing the psychometric quality of rating data. Psychological Bulletin. 88 (2), 413-428.

Wesdorp, H. (1981). Evaluatie-technieken voor het moedertaalonderwijs. Den Haag: Staatsuitgeverij, RITP, SVO.

Wilkinson, A. & Stratta, L. (1969). The evaluation of spoken language. Educational Review, 21, 183-195.